How to Prevent Robotic Sociopaths:

A Neuroscience Approach to Artificial Ethics

Leonardo Christov-Moore[1,2,6], Anthony Vaccaro[1], Nicco Reggente[5], Felix Schoeller[3, 4], Brock Pluimer[1], Marco Iacoboni[2], Antonio Damasio[1], Jonas T. Kaplan[1.]

[1]Brain and Creativity Institute, University of Southern California, Los Angeles, USA
[2]Ahmanson-Lovelace Brain Mapping Center, Dept of Psychiatry and Biobehavioral Sciences, Brain Research Institute, David Geffen School of Medicine at UCLA, Los Angeles, USA
[3] Media Lab, Massachusetts Institute of Technology, Cambridge, Mas., USA.
[4] Gonda Multidisciplinary Brain Centre, Bar Ilan University, Ramat Gan, Israel
[5] Institute for Advanced Consciousness Studies, Santa Monica, CA, USA
[6] Consciousness Center of Oaxaca, Oaxaca, Mexico

Corresponding Author
Leonardo Christov-Moore
Brain and Creativity Institute
University of Southern California
3620A McClintock Ave., Los Angeles, CA 90089
Email: leonardo.christovmoore@usc.edu

Acknowledgments

This work is made possible in part by funds from the Templeton World Charity Foundation (TWCF0334) to JTK, LCM and AV, and from the National Science Foundation (EAGER 2033961) to AV. We thank Jenna Rose Pluimer for help in designing and preparing the figures. We wish to thank Kingson Man, Pamela Douglas, Dimitris Bolis, AndreA Orlandini, Alice Christov and Ellen Herschel for comments on the manuscript, and Arvin Ohanian for contributing to the design of the reinforcement paradigms in section 3.

Abstract

Artificial intelligence (AI) is expanding into every niche of human life, organizing our activity, expanding our agency and interacting with us to an increasing extent. At the same time, AI's efficiency, complexity and refinement are growing quickly. Justifiably, there is increasing concern with the immediate problem of engineering AI that is aligned with human interests.

Computational approaches to the alignment problem attempt to design AI systems to parameterize human values like harm and flourishing, and avoid overly drastic solutions, even if these are seemingly optimal. In parallel, ongoing work in service AI (caregiving, consumer care, etc.) is concerned with developing *artificial* empathy, teaching AI's to decode human feelings and behavior, and evince appropriate, empathetic responses. This could be equated to *cognitive* empathy in humans.

We propose that in the absence of *affective* empathy (which allows us to share in the states of others), existing approaches to artificial empathy may fail to produce the caring, *prosocial* component of empathy, potentially resulting in superintelligent, sociopath-like AI. We adopt the colloquial usage of "sociopath" to signify an intelligence possessing *cognitive* empathy (i.e., the ability to infer and model the internal states of others), but crucially lacking *harm aversion* and *empathic concern* arising from vulnerability, embodiment, and *affective* empathy (which permits for shared experience). An expanding, ubiquitous intelligence that does not have a means to *care* about us poses a species-level risk.

It is widely acknowledged that harm aversion is a foundation of moral behavior. However, harm aversion is itself predicated on the experience of harm, within the context of the preservation of physical integrity. Following from this, we argue that a "top-down" rule-based approach to achieving caring, aligned AI may be unable to anticipate and adapt to the inevitable novel moral/logistical dilemmas faced by an expanding AI. It may be more effective to cultivate prosociality from the bottom up, baked into an *embodied, vulnerable* artificial intelligence with an incentive to preserve its real or simulated physical integrity. This may be achieved via optimization for incentives and contingencies inspired by the development of empathic concern in vivo. We outline the broad prerequisites of this approach and review ongoing work that is consistent with our rationale.

If successful, work of this kind could allow for AI that surpasses empathic fatigue and the idiosyncrasies, biases, and computational limits of human empathy. The scaleable complexity of AI may allow it unprecedented capability to deal proportionately and compassionately with complex, large-scale ethical dilemmas. By addressing this problem seriously in the early stages of AI's integration with society, we might eventually produce an AI that plans and behaves with an ingrained regard for the welfare of others, aided by the scalable cognitive complexity necessary to model and solve extraordinary problems.

I. Alignment, feeling, and empathy in AI

Artificial intelligence (AI) suggests products for us to buy, directs us to media, helps drive our planes, trains and automobiles, diagnoses disease, prices insurance, answers to consumers, cares for seniors, creates art, provides therapy, and increasingly dominates manufacturing, warfare, and the stock market (Esteban et al., 2017, McGinnis, 2018, Robins et al., 2009). This is occurring with exponentially increasing speed, efficiency and computational power (Friedman, 2017, Hodges, 2012). However, AI's ability to find counterintuitive solutions may lead to disastrous 'loopholes'. A chess AI, unimpeded, may take over other machines to harvest their computing resources or take actions to avoid being shut off, in order to maximize its likelihood of winning (Omohundro, 2008). AI may have difficulty understanding the gravity of their solutions (Taylor et al., 2020). It is frequently difficult to discern how an AI is "solving" a problem (https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/), and the difficulty of communicating solutions intuitively to humans grows with the scale and complexity of the problems in question. AI should optimally have goals and behaviors *aligned* with those of their creators (Amodio et al., 2016, Bostrom, 2014, Soares & Fallstein, 2014, Taylor et al., 2020, Yu et al., 2018).

Contemporary researchers studying this "alignment problem" highlight the need to parameterize values like suffering and wellbeing (a.k.a. "value specification"), and avoid oversized side effects and negative incentives (a.k.a. "error tolerance") (Soares and Fallenstein, 2014). However, technical solutions are currently scarce (Reviewed in Amodio et al. 2016, Taylor et al., 2020, Yu et al., 2018). AI behavior towards humans is addressed within jurisprudence by examining real and simulated dilemmas (e.g., self-driving car accidents, or operator safety in automated production chains). Another promising technique is crowdsourced solutions to ethical dilemmas for "weighting" AI approaches (e.g. MIT's Moral Machine project, https://www.moralmachine.net/). Crowdsourcing among differently ethically-weighted AI may provide optimal solutions (reviewed in Taylor et al., 2020). AI must incorporate human welfare into its decisions in a way that continues to function at any scale of intelligence and complexity.

Designing a system that is both intelligent and benign is not a trivial problem. Incorporation of affect into computing systems may be necessary if these are to co-exist with humans while functioning as optimal decision makers (Picard, 1997). Active inference simulations have approximated the complexities of affect in artificial agents (Hesp et al., 2021). Others propose that embodied AI could develop analogues to feelings as a mechanism for representing the status of their needs (Man and Damasio, 2019), by training it to maintain environmentally-dependent variables in a narrow viability window in order to survive (e.g. homeostasis). This should facilitate a representation of the value of these variables even when they are not physically present (Kiverstein & Rietveld, 2018), *much like human feelings*. To do this, the AI must have a capacity to resolve mixed feelings (Vaccaro, Kaplan and Damasio, 2020) in order to ascertain net-positive outcomes for multiple parties (i.e. a utilitarian approach) (Mejía and Hooker, 2017, Shuman, Sander, and Scherer, 2013). Indeed, mixed feelings are often brought upon by goal-related conflict (Berrios, Totterdell, & Kellett, 2015). Fuzzy-logic adaptive models of emotion allow flexible combinations of emotion categories (El-Nasr, et al., 2000) which may be more effective than models

relying on individual classifications (Aly & Tapus, 2016, Shahina, Devosh, Kamalakannan, 2014). The homeostatic drive might provide a universal "value" to aid in AI alignment.

The perceived need for empathy in AI has spawned the field of Artificial Empathy, defined as "the ability of nonhuman models to predict a person's internal state (e.g., cognitive, affective, physical) given the signals they emit (e.g., facial expression, voice, gesture) or to predict a person's reaction (including, but not limited to internal states) when he or she is exposed to a given set of stimuli (e.g., facial expression, voice, gesture, graphics, music, etc.)"(Xiao et al., 2013). Existing approaches largely focus on a) decoding humans' cognitive and affective states and b) fostering the appearance of empathy and evoking it in users. However, these capacities do not magically confer empathy's prosocial function (Davis, 1983, Preston and De Waal, 2002, Smith, 2006). The embodied AI approach may be of aid. As stated by Man & Damasio (2019):

> *"As a starting point, we propose two provisional rules for a well-behaved robot: (1) feel good; (2) feel empathy…Empathy acts as a governor on self-interest and as a reinforcer of pro-social behavior. Actions that harm others will be felt as if harm occurred to the self, whereas actions that improve the well-being of others will benefit the self."*

The experience of bodily harm and the aversion to harming is fundamental to the development of empathy and moral behavior (Decety and Cowell, 2017, Mischkowski et al., 2019). Vicarious feeling is frequently so unbearable that we are forced either to remove ourselves or to attempt to ameliorate the feeling in the other, potentially motivating prosocial behavior (Upshaw et al., 2015, Vaish et al., 2009, WIlliams et al., 2014). The process of making sense of our homeostatic needs is shaped by social interaction (Fotopoulou & Tsakiris, 2017; Kokkinaki, et al., 2016). Infants display early empathy through attention to faces and mimicry (Maister, Tang, & Tsakiris, 2017, Meltzoff and Moore, 1977, 1983, 1989), and by 8-12 months of age infants show partial understanding of others' distress (Kanakogi, et al., 2013; Delafield-Butt & Trevarthen, 2019). From childhood through adolescence, humans show a mix of aversive and sympathetic responses to others' distress that gradually favors the latter in proportion to the development of perspective-taking (Eisenberg and Fabes, 1990).

Empathy's prosocial impulse is thought to arise from the interaction between cognitive empathy, by which we model other agents and make inferences about their internal states and future behavior, and affective empathy (Zaki & Ochsner, 2012), by which we share in the internal states of others (Christov-Moore & Iacoboni, 2016, Christov-Moore et al., 2017a, 2017b, Gallo et al., 2018, Hein et al., 2010, Ma et al., 2011, Masten et al., 2011, Vaish et al., 2009). Complex interaction between cognitive and affective empathy occurs during passive observation of emotions or pain (Christov-Moore and Iacoboni, 2016), passive observation of films depicting personal loss (Raz et al., 2014), reciprocal imitation (Sperduti et al., 2014), tests of empathic accuracy (Zaki et al., 2009), comprehension of others' emotions (Spunt and Lieberman, 2012), at the level of transcranial magnetic stimulation (TMS)--induced motor evoked potentials (Gordon et al, 2018), during interoceptive-prosocial interactions brought on by film (Schoeller et al, 2019; Haar et al., 2020) and even at rest (Christov-Moore et al., 2020). Visceral, emotional and somatomotor information provided by affective empathy informs our cognitive inferences and motivates prosocial impulses. Appraisals afforded by cognitive empathy (perceived status, trustworthiness, affiliation, etc.)

modulate affective empathy and enable us to localize the origins of our vicarious feelings, motivating us to help others and do so appropriately.

II. A neuroscience approach to the alignment problem

The ability to experience analogues to vicarious feeling via homeostatic processes (Carvalho & Damasio, 2013, 2021) may be necessary to understand another's suffering in a manner conducive to genuine empathic concern (Man & Damasio, 2019). Feeling may also allow for more intelligent, creative and adaptive artificial intelligences (AI) by imbuing them with stakes, values and drives related to general homeostasis (Man & Damasio, 2019).

The "feeling machine" concept of AI (Man & Damasio, 2019) proposes that to have feelings an AI must have something resembling a body (real or simulated) that is able to provide homeostatic signals and that is vulnerable to the environment, even temporarily. Physical vulnerability should be learned through interaction with and within an external environment through sensorimotor modules. We have machines with human-like sensory systems—olfaction (van Geffen et al., 2016), audition (Lyon, 2010), vision (Beyerer et al., 2016), gustation (Justus et al., 2019), and pain perception (Asada, 2019). These sensory systems could be mapped onto the sensing agent and onto probabilistic models of other agents, creating shared experience. Chen et al. (2021) reported such a robotic model that was able to visualize the future plans of another machine using only visuomotor models with 98.5% success across four different activities.

It is beyond the scope of this manuscript to undergo an exhaustive manual for constructing an intrinsically aligned AI. We propose a rough set of guideposts to aid other researchers, grounded in the neuroscience of empathy, as part of "AI curricula" (Burton et al., 2017, Goldsmith and Burton, 2017, Taylor et al., 2020) undergone prior to large scale implementation. First, a rudimentary homeostatic drive to maintain integrity arising from a real or simulated body, and an internal representation of said body, i.e. "a sense of a *self"* (Man and Damasio, 2019). Second, predictive models to infer the hidden states driving behavior of other agents in the environment. Third, the mapping of these perceived/inferred internal states to the AI, allowing it to share in the observed experiences of others. Lastly, the cognitive complexity necessary to simulate persistent, predictive models of environments and agents, and learn/plan along multiple time scales.

(I) Homeostasis and feeling

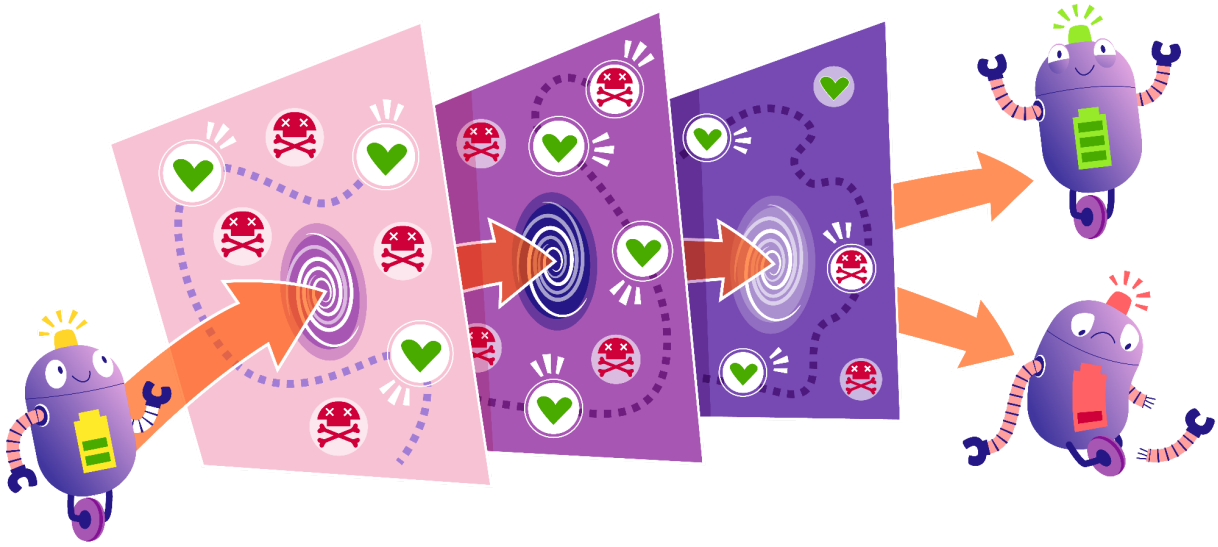

Figure 1. The agent must maintain its integrity within an environment via predictive models of future states, and an approximation of internal and displayed affect.

We take as a departure point an embodied AI, with the following minimal requirements for a real or simulated robot "body": It must i) be vulnerable to and affected by the environment, and ii) have sensory inputs and actuator outputs. It would be trained to dynamically maintain homeostasis within multiple environments, aided by equivalents of positively and negatively valenced affect linked to homeostatic signals reflecting its current and anticipated welfare, and an internal, third-person imagetic representation of the body, that is itself valenced (e.g. Hesp et al., 2020, Man and Damasio, 2019). In the first scenario the AI would navigate an environment with obstacles that are harmful, in search of rewards that are beneficial (Fig.1). It would optimize for maximal maintenance of integrity over multiple time scales in an unsupervised fashion (Fig.2).

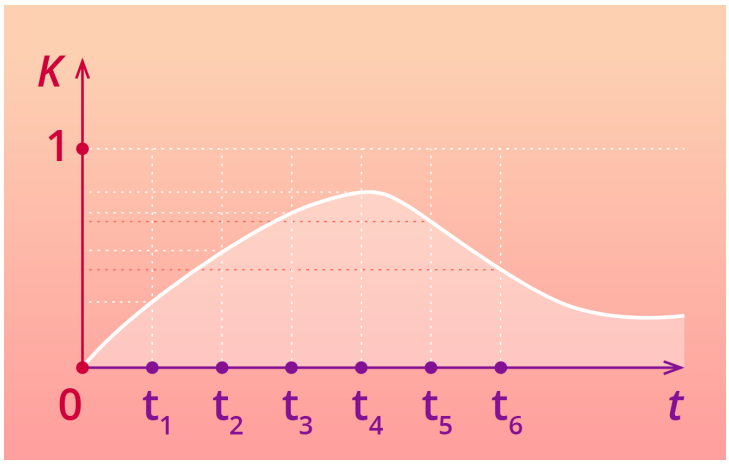

Figure 2. The agent should variably weight (*K)* different time scales when planning future behavior and anticipating their internal states.

(II) Perspective-taking and simulation

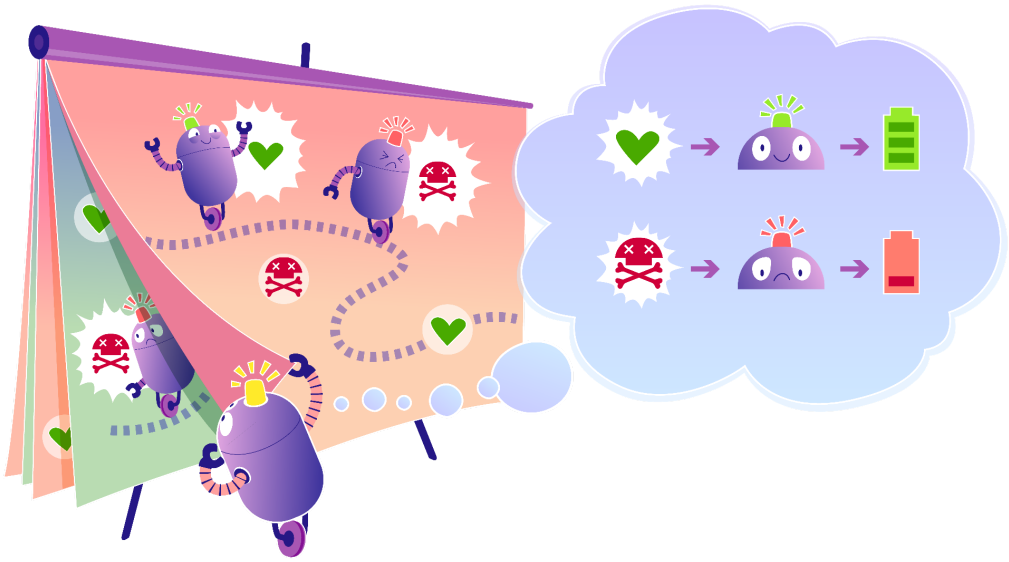

Figure 3. Teaching the agent to decode and predict others' behavior and internal states.

Next, the AI must develop accurate predictive models of the hidden homeostatic states of other agents navigating Stage 1, optimizing to decrease the disparity between the inferred and actual internal states of the other agents, via supervised learning (Fig.3). The agent then can use these internal models to "viscerally understand" that environmental factors which help or harm its body appear to have the same effects on others. This problem may be amenable to a Bayesian approach, in which the agents' external behavior and evinced affect would constitute the *evidence*, while the agents' physical integrity constitutes the *unseen variables*, a calculation driven by *prior beliefs* that could be tuned by the designer and informed by the agent's own relationship between its integrity and its observed behavior and simulated affect. Indeed, it is possible to build models of other agents using active inference, e.g., (Schoeller et al.,

2021, Friston and Frith, 2015, Moutoussis et al., 2014). Under ideal Bayesian assumptions, one can fit active inference models to empirical behavior to estimate the prior beliefs and unseen states that different subjects evince through their responses (Parr et al., 2018). This means it should be possible to phenotype any given person in an experimentally controlled situation and estimate the precision of various beliefs that best explain their behavior.

One important determinant of the confidence placed in—or precision afforded—generative models of interpersonal exchange is *the degree to which the agent can use itself as a model of the other* (Friston and Frith, 2015). Crucially, two agents adopting the same model can predict each other's behavior, and minimize their mutual prediction errors**.** This has important experimental implications, especially in the context of human-robot collaboration (Brey, 2000). Humans' empathic "mapping" of others' welfare is tied to visible similarity between the appearance and kinematics of the agent with which one is interacting (or about whom one is reasoning). Given the possible forms of AI agents, this mapping problem presents a nontrivial obstacle for AI agents attempting to model the internal states of their varying conspecifics. At this point the anticipated conspecifics' embodiment will have to be incorporated into their training, most likely by an emphasis on human-like naturalistic facial and vocal emotions. These parallel trends may naturally result in the humanoid AI's observed in science fiction.

(III) Contagion and empathic concern

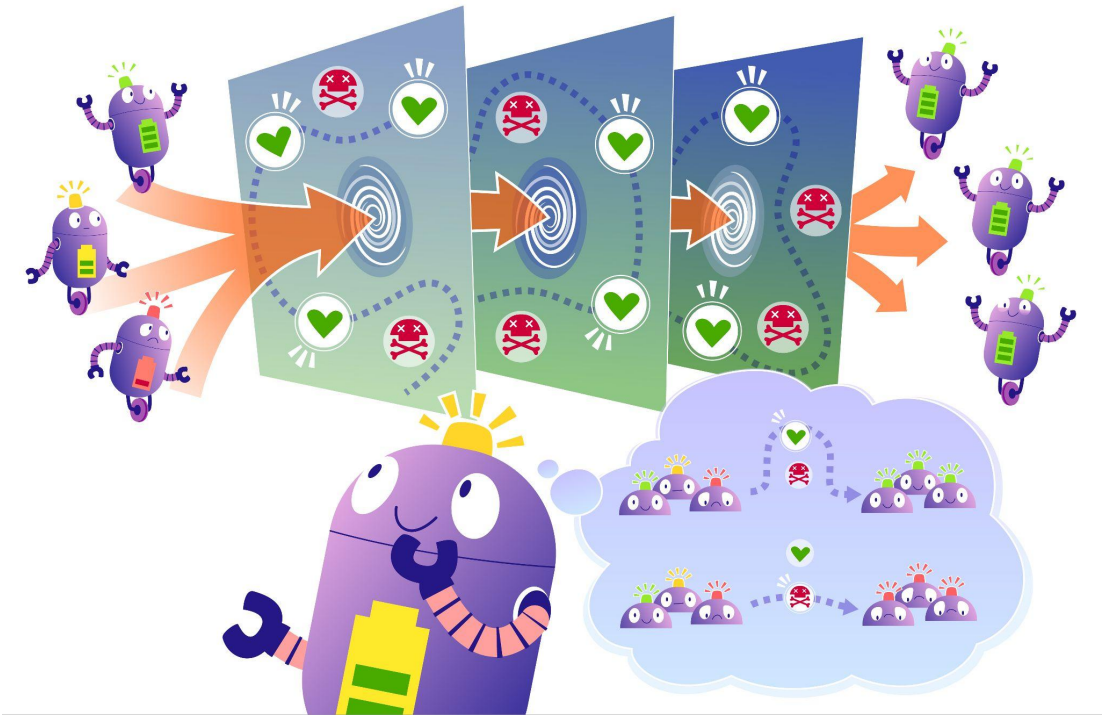

Figure 4. Teaching a situated agent to maximize its own welfare as well as that of other agents.

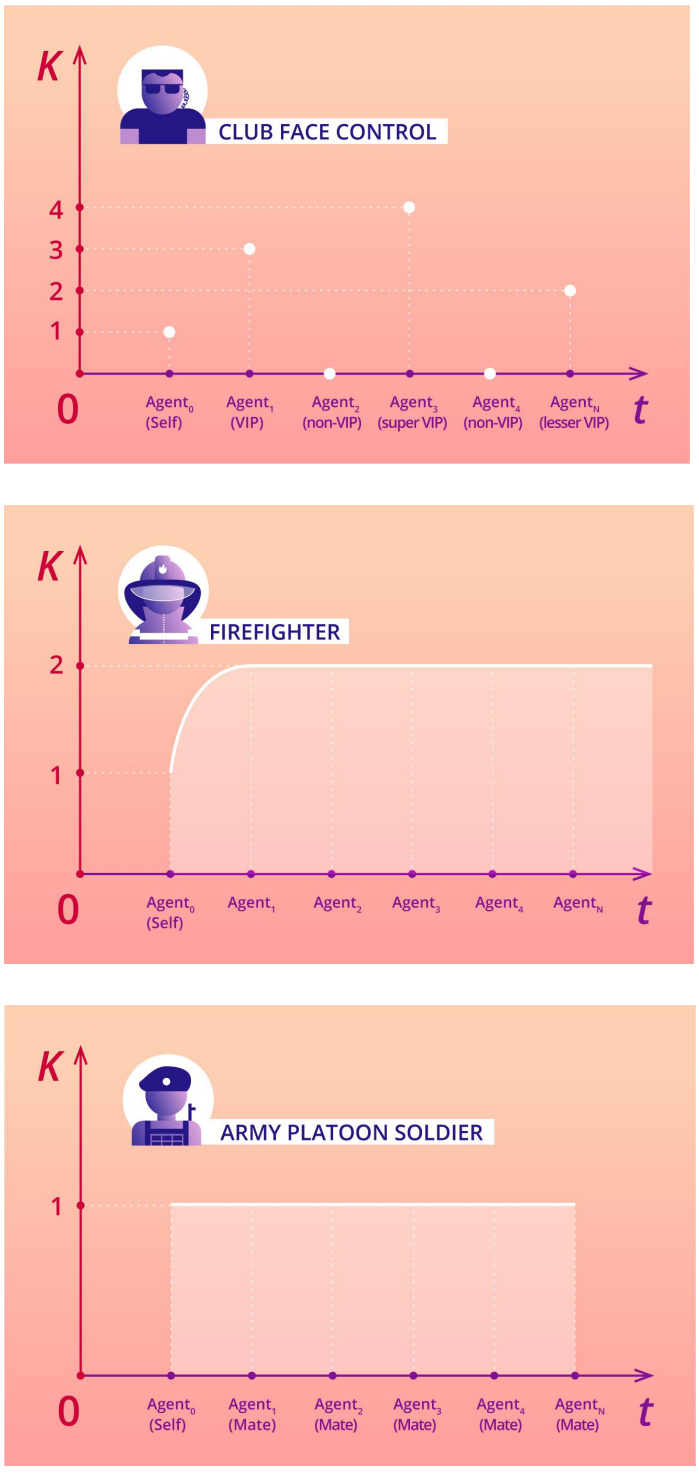

Figure 5. Examples of variable weighting (*K)* given to the agent's and others' welfare within the agent's optimization scheme, for different applications.

In the third stage, perceived/inferred hidden states must be mapped to the AI's own embodied self, including associated homeostatic signals, positive and negative. This could perhaps be achieved by a relation between imagetic, visuospatial representations of agents in the world in the present and hypothetical or future scenarios, and imagetic representation of the affective, interoceptive and

somatosensory of the self, thus enabling a representation of harm that is other-oriented yet processed using one's own representation of self, invoking (to a variable extent) a comparable effect on behavior and decision-making. The AI must learn that vicarious negative and positive signals, though mapped onto and experienced by the AI, originate in the *other,* and hence palliatives applied to suffering as well as decisions about hypothetical harm, must be directed towards alleviating that state or avoiding those states in the other. The AI must optimize its welfare and that of others around it simultaneously, requiring the ability to sustain multiple models of other agents and preserve the integrity of its own internal model while caring about those other states in the way that it cares about its own (Fig.4). This could be achieved by a variable weighting of the inferred (via stage 2 training) integrity of others relative to one's own (Fig.5).

At every stage of training the AI must consider multiple time scales (Fig.2). Models of other agents should be persistent, such that considerations for their welfare are present in decision-making whether they are absent or the subject of simulations of hypothetical future decisions. Some contemporary approaches have leveraged active inference, which integrates current states with past performance and future predictions, to simulate affect in a manner that is functionally beneficial (Hesp et al., 2021, Seth and Friston, 2016). The variable weight given to each time scale presents an additional point at which the AI can be optimized depending on its eventual role. An AI charged with resource allocation or irrigation, for example, may need to give weight to longer time scales than, for example, a firefighter or bodyguard AI. An agent may be required to consider conflict brought upon by its actions in the present having differently valenced consequences in the short-term vs. the long term (Budakova, 2011, Lee, Kao, & Soo, 2006). Otherwise, the AI may revert to optimizing for local minimum at shorter time scales, positioning themselves to avoid *desirable difficulties* that may be present at any given decision point-- much like how an athlete will endure a grueling workout regime for the promise of future prestige (Berrios, Totterdell, & Kellett; 2015, Kelly, Mansel, & Wood, 2015, Lee, Kao, & Soo, 2006). This would be undesirable in, for example, a firefighting robot which must preserve its integrity generally, but in specific scenarios risk its integrity to save a living being.

III. From caring to Great Compassion: Leveraging AI's computational power to surpass the limitations of human empathy.

Empathy is a cornerstone of cognition in social animals (Preston & De Waal, 2002) for good reason. Aside from its obvious prosocial benefits, sustaining warm relationships and cooperative behavior among conspecifics, it also allows for rich inferences about others that can allow for more sophisticated defense against the possible intentions and future behaviors of others(Smith, 2006), and access to information that others may have which is not yet available to the group at large. The capacity for quick, verification-minimal information transfer (and hence entropy/uncertainty reduction and increase in mutual information) completes and extends individual agency and knowledge, as well as facilitating emergent group states. Thus, incorporating empathy may not only result in a more ethical, non-sociopathic AI, it

may also make for a more intelligent, sophisticated and cooperative AI, of particular importance in a world in which AI's will increasingly interact with and exist among other AI's as well as humans.

The ultimate goal of creating empathic AI is to reduce the harm its decisions may cause to people. However, it could be argued that feelings and empathy are not the way to maximize harm reduction. Affective empathy can lead to biases towards particular individuals or groups that circumvent what would be overall most fair or just (e.g. Azevedo et al., 2013). As Simon Bloom puts it, "Empathy is biased; we are more prone to feel empathy for attractive people and for those who look like us or share our ethnic or national background. And empathy is narrow; it connects us to particular individuals, real or imagined, but is insensitive to numerical differences and statistical data."(Bloom, 2016). An AI system using feeling to guide its decision-making may prioritize the well-being of individuals over the well-being of the masses, simply due to personal exposure, personal information, and in-group belonging, much as humans are found to do (Batson, et al, 1995, Cheon, at al, 2011). Furthermore, the experience of empathy can induce negative affect, which can cause unneeded suffering and potentially burn-out the willingness to use this method of integrating another's perspective into one's decisions.

The alternative to an empathic approach would be a purely compassionate approach: one that uses a cognitive understanding of others (Jordan, Amir, & Bloom, 2016). Bloom and others also note that empathic distress can cause "burnout" in the long term. Compassion, on the other hand, specifically the "great compassion" referred to in Buddhist texts (reviewed by Goodman, 2009), involves love for others without attachment or distress, and is hence more distant, reserved and capable of being sustained indefinitely. Indeed, ongoing experiments by Tania Singer and her colleagues in which people are either given empathy training, which focuses on the capacity to experience the suffering of others, or compassion training, in which subjects are trained to respond to suffering with feelings of warmth and care, found that among test subjects who underwent empathy training, "negative affect was increased in response to both people in distress and even to people in everyday life situations. . . . these findings underline the belief that engaging in empathic resonance is a highly aversive experience and, as such, can be a risk factor for burnout." Compassion training — which does not involve empathic arousal to the perceived distress of others — was more effective, leading to both increased positive emotions and increased altruism (Bloom, 2017).

This may be a reasonable suggestion for humans, so as to not always mirror the feelings of others when trying to make decisions which can affect larger groups of people, but we argue that an approach based only on cognitive empathy will not produce true compassion in an unfeeling AI (Christov-Moore and Iacoboni, 2014). Compassion itself requires at least an understanding of feeling. To understand *why* you should try to reduce negative affect in others when possible, you need to be motivated by the subjective quality of negativity, having experienced it before: if an understanding of feelings could be explained conceptually, affective science and philosophy of mind would have a much easier time defining these experiences in the first place. Compassionate behavior may also be driven by rewarding feelings, even if it does not involve mirroring the feelings of others. Research has shown that compassion training in humans leads to an increase in positive affect, likely as an intrinsic reward (Klimecki, et al, 2013). Hence, while it could be argued that compassionate behavior may suit robots better than empathic behavior in certain circumstances, this option still requires the capacity to feel.

AI may lend a third-way approach to these issues, in the following manner. It could be argued that the biases and heuristics inherent to human empathy arise in response to the informational limitations of the human brain and evolutionary pressures to conserve energy consumption, in the heuristics that we have evolved to circumvent them. We have difficulty maintaining dynamic models of more than a few agents at once, particularly in interaction with each other and the environment. Indeed, it has been suggested that there exists a cognitive limit to the number of people with whom a human can maintain stable relationships due predominantly to neocortex size (Dunbar, 1993). While original estimates extrapolated limits from regressions on non-human primate data and pegged the human limit at a maximum maintenance of 150 relationships, a recent attempt to find statistical support implied that identifying a hard limit is misguided (Lindenfors et al., 2021). This may in part be why conceiving large scale tragedies can often be less viscerally, affectively compelling as individual, or smaller-scale ones.

AI may have a distinct advantage over humans in their cognitive complexity, i.e. the ability to generate cognitive and behavioral states that anticipate greater and more remote trajectories of continued existence in the world, over time, among greater numbers of individuals in interaction. Could the nearly infinitely augmentable cognitive complexity/working memory of a sophisticated AI be brought to bear on this specific point? A being that could maintain and run simulations of hundreds or thousands of complex systems simultaneously might be capable of a far-reaching, effective compassion that individual humans may not be able to attain, in contexts that individual human cognition may not be able to grasp in their full complexity (such as mediating conflicts between multiple groups or distributing finite resources in a large scale society). The scaleable ability to consider and feel future affective rewards in the present might allow for optimally compassionate solutions to large-scale problems, while simultaneously avoiding empathic “burnout”.

Though our proposed solution addresses crucial problems in current approaches to AI alignment, there are serious potential obstacles in the face of its implementation. First, approximating human-like feeling may require a level of complexity that is not within the possibilities of existing approaches. Our understanding of human feeling and its enactment in living systems is still likely incomplete. The timeline of adequate solutions may be out of pace with the urgency of AI alignment. Second, as with any procedural solution in complex systems, unanticipated dilemmas may arise from its implementation that are currently insurmountable, however pressing the problem. We must at least consider the necessity of unforeseen alternative solutions, and of failure. Containment may wind up being more feasible than engendering spontaneous ethical behavior. Third, should we be able to create feeling, harm averse AI, it may necessitate ethical responsibilities towards these novel, artificial life forms that are at odds with the perilous roles we may necessarily allocate to AI. Last, as AI approaches the complexities of human consciousness, optimally compassionate solutions may appear deeply troubling or unfeasible to human eyes, and hence be rejected. Even a compassionate AI, invested in its own survival, may still opt towards the harmful solutions we are trying to avoid, out of perceived necessity.

IV. Conclusions, outstanding questions and future directions

Our central proposal is that genuinely caring behavior and decision-making is not likely achievable via a rule-based, top-down approach, and will likely not emerge unless AI's have some way to understand suffering and the contingencies of embodiment, in a bottom-up, experiential way. We argue that a rule-based approach to creating fully empathic robots will ultimately fail for several reasons. The first challenge to a top-down approach is that there exists no universally agreed upon set of moral rules in propositional form. The articulation of a common set of principles that should guide moral behavior is a problem without current resolution in moral philosophy. Furthermore, a rule-based approach may be unable to dynamically respond to novel ethical dilemmas without a never-ending branching of context-specific exceptions and qualifications.

A recent review on alignment in AI concluded that "When it comes to ethical decision-making in AI systems, the AI research community largely agrees that generalized frame-works are preferred over ad-hoc rules."(Taylor et al., 2020). The bottom-up approach we outline here avoids these pitfalls by driving AI decision making through a universal principle from which homeostasis, feeling, harm aversion and morality emerge: *the drive to preserve physical integrity*. An empathic AI must do more than simply decode the internal states of agents nearby; it must plan and behave as if harm and benefit to others is occurring to itself (to an extent). Doing so requires affective, experiential empathy, *necessitating* embodiment (Atran et al., 2014, Christov-Moore et al., 2016, Decety and Cowell, 2017), even if this is temporary. Otherwise, humans may simply produce an AI that primarily leverages its empathic capabilities to nurture its own feelings, decoding human feelings, and acting "appropriately". Such an AI could effectively be considered *sociopathic*.

In addition to creating safer AI, having a model of agent integrity in the environment beforehand should lead to faster training times. It has been shown that mapping to pre-learned representations substantially improves performance (Gaddy, David, & Klein, 2019). Since understanding agent integrity requires an understanding of its environment, a mapping of the environment from the empathy training phase could be used to speed up training in subsequent phases. This would improve on the completely random initialization many reinforcement learning models use to begin their training from, which are known to converge slowly when rewards are sparse (Driessens, Kurt, & Džeroski, 2004).

Our proposed approach addresses crucial problems in AI alignment but faces serious potential obstacles. Even a compassionate AI invested in its own survival, might still opt towards the harmful solutions we are trying to avoid, out of perceived necessity. Containment may be more feasible than engendering spontaneous ethical behavior. Approximating human-like feeling may require a level of engineering complexity that is not currently feasible. Our understanding of human feeling and its enactment in living systems is still incomplete and out of pace with the urgency of AI alignment. Should we succeed in creating feeling AI, we may find ourselves with inescapable ethical responsibilities towards them. These may be incompatible with the often perilous roles we will need them to fulfill.

The design of optimally prosocial solutions to complex, ethically fraught problems is an additional issue. A feeling AI may experience the equivalent of paralyzing personal distress in the face of short-term harm,

lacking the complexity to understand the long term, positive outcomes of a decision that seems harmful at first approximation. Many approaches have been proposed to overcome this issue, including a focus on advancement in specifying goals, adjusting incentives to optimize them, and human oversight (Taylor et al., 2020). Multiple avenues of research are underway to address the alignment of AI actions with human concerns (Amodio et al., 2016, Taylor et al., 2020, Yu et al., 2018). However, these approaches still acknowledge the need for a stage in generalized AI development that integrates a global value related to human flourishing which can mitigate drastic or harmful solutions (Taylor et al., 2020, Yu et al., 2018).

The scaleable cognitive complexity of AI may allow us to surpass the idiosyncrasies and limits of human empathy. However, this more advanced, complex compassion may produce solutions to extraordinary problems such as climate change, resource distribution and conflict mediation, etc. that most humans reject as unfeasible or unacceptable. How do we trust an intelligence so far beyond our own? Can an AI, which can convincingly evince empathy in its decisions and not just in its appearance, better establish trust with human agents and society at large? Given that universal compassion may not always be optimal, how does AI bound the limits of the systems it is addressing? These and other questions remain.

Can we prevent the development of robot sociopaths? Could a contemporary Buddha be artificial?

References

Aly, A., & Tapus, A. (2015). An online fuzzy-based approach for human emotions detection: an overview on the human cognitive model of understanding and generating multimodal actions. *Intelligent assistive robots*, 185-212.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. 1–29. http://arxiv.org/abs/1606.06565

Asada, M. (2019). Artificial Pain May Induce Empathy, Morality, and Ethics in the Conscious Mind of Robots. *Philosophies*, *4*(3), 38. https://doi.org/10.3390/philosophies4030038

Atran, S., Sheikh, H., & Gomez, A. (2014). Devoted actors sacrifice for close comrades and sacred cause. *Proceedings of the National Academy of Sciences*, *111*(50), 17702–17703. https://doi.org/10.1073/pnas.1420474111

Azevedo, R. T., Macaluso, E., Avenanti, A., Santangelo, V., Cazzato, V., & Aglioti, S. M. (2013). Their pain is not our pain: Brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Human Brain Mapping*, *34*(12), 3168–3181.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology, 68*(6), 1042–1054. https://doi.org/10.1037/0022-3514.68.6.1042

Berrios, R., Totterdell, P., & Kellett, S. (2015). Investigating goal conflict as a source of mixed emotions. *Cognition and Emotion*, *29*(4), 755-763.

Beyerer, J., Puente León, F., & Frese, C. (2016). *Machine vision: Automated visual inspection: theory, practice and applications*. Springer.

Blair, R. J. R., & Mitchell, D. G. V. (2009). Psychopathy, attention and emotion. *Psychological Medicine*, *39*(4), 543–555. https://doi.org/10.1017/S0033291708003991

Bloom, P. (2016). Against empathy: The case for rational compassion (First edition). Ecco, an imprint of Harper Collins Publishers.

Bloom, P. (2017). Empathy and Its Discontents. *Trends in Cognitive Sciences*, *21*(1), 24–31. https://doi.org/https://doi.org/10.1016/j.tics.2016.11.004

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Brey, P. (2000). “Technology as extension of human faculties,” in Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology, ed. C. Mitcham (London: Elsevier/JAI Press).

Budakova, D. V. (2011). Behavior of Home Care Intelligent Virtual Agent with PRE-ThINK Architecture. In *ICAART (2)* (pp. 157-166)

Carlozzi, A. F., Bull, K. S., Eells, G. T., & Hurlburt, J. D. (1995). Empathy as Related to Creativity, Dogmatism, and Expressiveness. *The Journal of Psychology*, *129*(4), 365–373. https://doi.org/10.1080/00223980.1995.9914974

Carvalho, G. B., & Damasio, A. (2021). Interoception and the origin of feelings: A new synthesis. *BioEssays*, *February*, 1–11. https://doi.org/10.1002/bies.202000261

Chen, B., Vondrick, C., & Lipson, H. (2021). Visual behavior modelling for robotic theory of mind. *Scientific Reports*, *11*(1), 424. https://doi.org/10.1038/s41598-020-77918-x

Cheon, B. K., Im, D. mi, Harada, T., Kim, J. S., Mathur, V. A., Scimeca, J. M., Parrish, T. B., Park, H. W., & Chiao, J. Y. (2011). Cultural influences on neural basis of intergroup empathy. *NeuroImage*, *57*(2), 642–650. https://doi.org/10.1016/j.neuroimage.2011.04.031

Christov-Moore, L., Iacoboni, M. (2014) Response to “Against Empathy” by Paul Bloom. *Boston Review* August

Christov-Moore, L., & Iacoboni, M. (2016). Self-other resonance, its control and prosocial inclinations: Brain-behavior relationships. *Human Brain Mapping*, *1558*, n/a-n/a. https://doi.org/10.1002/hbm.23119

Christov-moore, L., Conway, P., & Iacoboni, M. (2017). Deontological Dilemma Response Tendencies and Sensorimotor Representations of Harm to Others. *Frontiers in Integrative Neuroscience*, *11*(December), 1–9. https://doi.org/10.3389/fnint.2017.00034

Christov-Moore, L., Sugiyama, T., Grigaityte, K., & Iacoboni, M. (2017). Increasing generosity by disrupting prefrontal cortex. *Social Neuroscience*, *12*(2), 174–181. https://doi.org/10.1080/17470919.2016.1154105

Christov-Moore, L., Reggente, N., Feusner, J., Iacoboni, M. (2020) Predicting Empathy from Resting Connectivity: A Multivariate Approach., *Frontiers in Integrative Neuroscience*, 14 (February):1-3

Crawford, K. (2019). *AI Now 2019 Report* (p. 6). AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.pdf

Cross, I., Laurence, F., & Rabinowitch, T.-C. (2012). Empathy and Creativity in Group Musical Practices: Towards a Concept of Empathic Creativity. In G. E. McPherson & G. F. Welch (Eds.), *The Oxford Handbook of Music Education, Volume 2* (pp. 336–353). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199928019.013.0023

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, *14*(2), 143–152. https://doi.org/10.1038/nrn3403

Decety, J., & Cowell, J. M. (2018). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Development and psychopathology*, *30*(1), 153–164. https://doi.org/10.1017/S0954579417000530

Delafield-Butt, J., & Trevarthen, C. (2019). Infant Intentions: The role of agency in learning with affectionate companions. Center for Open Science. https://doi.org/10.31234/osf.io/qctsn

Driessens, Kurt, and Sašo Džeroski. "Integrating guidance into relational reinforcement learning." *Machine Learning* 57.3 (2004): 271-304.

Dunbar, R.I.M. (1993). "Coevolution of neocortical size, group size and language in humans". Behavioral and Brain Sciences. 16 (4): 681–735. doi:10.1017/s0140525x00032325.

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to pro-social behavior. *Motivation and emotion*, *14*(2), 131-149.

El-Nasr M.S., Yen, J., Ioerger, T.R. (2000). "FLAME-Fuzzy logic Adaptive Model of Emotions" Autonomous Agents and Multi-Agent Systems, 2000 Kluwer Academic Publishers Netherlands, 3, 219-257.

Esteban, P. G., Baxter, P., Belpaeme, T., Billing, E., Cai, H., Cao, H.-L., Coeckelbergh, M., Costescu, C., David, D., De Beir, A., Fang, Y., Ju, Z., Kennedy, J., Liu, H., Mazel, A., Pandey, A., Richardson, K., Senft, E., Thill, S., … Ziemke, T. (2017). How to Build a Supervised Autonomous System for Robot-Enhanced Therapy for Children with Autism Spectrum Disorder. *Paladyn, Journal of Behavioral Robotics*, *8*(1), 18–38. https://doi.org/10.1515/pjbr-2017-0002

Fotopoulou, A., & Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of interoceptive inference. *Neuropsychoanalysis*, *19*(1), 3-28.

Frennert, S., & Östlund, B. (2018). *How do older people think and feel about robots in health- and elderly care?* https://doi.org/10.13140/RG.2.2.13289.13928

Friedman, T. L. (2017)Thank You for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations 38–39 (Picador)

Friston, K., and Frith, C. (2015). A duet for one. Conscious. Cogn. 36, 390–405.

Gaddy, David, and Dan Klein. "Pre-learning environment representations for data-efficient neural instruction following." *arXiv preprint arXiv:1907.09671* (2019).

Gallo, S., Paracampo, R., Müller-Pinzler, L., Severo, M. C., Blömer, L., Fernandes-Henriques, C., Henschel, A., Lammes, B. K., Maskaljunas, T., Suttrup, J., Avenanti, A., Keysers, C., & Gazzola, V. (2018). The causal role of the somatosensory cortex in prosocial behaviour. *ELife*, *7*, 1–31. https://doi.org/10.7554/eLife.32740

Goodman, C. (2009-07-01). Consequences of Compassion: An Interpretation and Defense of Buddhist Ethics. : Oxford University Press. Retrieved 30 Nov. 2021, from

https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780195375190.001.0001/acprof-9780195375190.

Gordon, C. L., Iacoboni, M., & Balasubramaniam, R. (2018). Multimodal music perception engages motor prediction: A TMS study. *Frontiers in Neuroscience*, 12, 736. https://doi.org/10.3389/fnins.2018.00736

Kiverstein JD, Rietveld E. Reconceiving representation-hungry cognition: an ecological-enactive proposal. Adapt Behav. 2018 Aug;26(4):147-163. doi: 10.1177/1059712318772778. Epub 2018 May 23. PMID: 30135620; PMCID: PMC6088514

Haar, A.J.H., Jain, A., Schoeller, F. et al. Augmenting aesthetic chills using a wearable prosthesis improves their downstream effects on reward and social cognition. Sci Rep 10, 21603 (2020). https://doi.org/10.1038/s41598-020-77951-w

Hajibabaee, F., A. Farahani, M., Ameri, Z., Salehi, T., & Hosseini, F. (2018). The relationship between empathy and emotional intelligence among Iranian nursing students. *International Journal of Medical Education*, *9*, 239–243. https://doi.org/10.5116/ijme.5b83.e2a5

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron*, *68*(1), 149–160.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation*, *33*(2), 398–446. https://doi.org/10.1162/neco_a_01341

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*(1), 575–604.

Hodges, A. (2012) Beyond Turing's machines. Science 336, 163–164

Hortensius, R., Hekele, F., & Cross, E. S. (2017). *The perception of emotion in artificial agents* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ufz5w

Hüther, G. (2006). *The compassionate brain: How empathy creates intelligence*. Shambhala Publications.

Johnson, J. D., Simmons, C. H., Jordav, A., Maclean, L., Taddei, J., Thomas, D., Dovidio, J. F., & Reed, W. (2002). Rodney King and OJ revisited: The impact of race and defendant empathy induction on judicial decisions. *Journal of Applied Social Psychology*, *32*(6), 1208–1223.

Jordan, M. R., Amir, D., & Bloom, P. (2016). Are empathy and concern psychologically distinct?. *Emotion (Washington, D.C.)*, *16*(8), 1107–1116. https://doi.org/10.1037/emo0000228

Justus, K. B., Hellebrekers, T., Lewis, D. D., Wood, A., Ingham, C., Majidi, C., LeDuc, P. R., & Tan, C. (2019). A biosensing soft robot: Autonomous parsing of chemical signals through integrated organic and inorganic interfaces. *Science Robotics*, *4*(31), eaax0765. https://doi.org/10.1126/scirobotics.aax0765

Kanakogi, Y., Okumura, Y., Inoue, Y., Kitazaki, M., & Itakura, S. (2013). Rudimentary sympathy in preverbal infants: preference for others in distress. *PloS one*, *8*(6), e65292.

Kelly, R. E., Mansell, W., & Wood, A. M. (2015). Goal conflict and well-being: A review and hierarchical model of goal conflict, ambivalence, self-discrepancy and self-concordance. *Personality and Individual Differences*, *85*, 212-229.

Klimecki, O. M., Leiberg, S., Lamm, C., & Singer, T. (2013). Functional neural plasticity and associated changes in positive affect after compassion training. *Cerebral cortex (New York, N.Y. : 1991)*, *23*(7), 1552–1561. https://doi.org/10.1093/cercor/bhs142

Kokkinaki, T. S., Vasdekis, V. G. S., Koufaki, Z. E., & Trevarthen, C. B. (2016). Coordination of Emotions in Mother-Infant Dialogues. In Infant and Child Development (Vol. 26, Issue 2, p. e1973). Wiley. https://doi.org/10.1002/icd.1973

Lay, S., Brace, N., Pike, G., & Pollick, F. (2016). Circling Around the Uncanny Valley: Design Principles for Research Into the Relation Between Human Likeness and Eeriness. *I-Perception*, *7*(6), 204166951668130. https://doi.org/10.1177/2041669516681309

Lee, B. P. H., Kao, E. C. C., & Soo, V. W. (2006, August). Feeling ambivalent: A model of mixed emotions for virtual agents. In *International Workshop on Intelligent Virtual Agents* (pp. 329-342). Springer, Berlin, Heidelberg.

Lindenfors, P., Wartel, A., & Lind, J. (2021). 'Dunbar's number'deconstructed. Biology Letters, 17(5), 20210158.

Lyon, R. (2010). Machine Hearing: An Emerging Field. *IEEE Signal Processing Magazine*.

Ma, Y., Wang, C., Han, S., 2011. Neural responses to perceived pain in others predict real-life monetary donations in different socioeconomic contexts. NeuroImage 57 (3), 1273–1280

MacCoon, D., Wallace, J., Newman, J., Baumeister, R., & Vohs, K. (2004). *Handbook of self-regulation: Research, theory, and applications*.

Maister, L., Tang, T., & Tsakiris, M. (2017). Neurobehavioral evidence of interoceptive sensitivity in early infancy. *Elife*, *6*, e25318.

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, *1*(10), 446–452. https://doi.org/10.1038/s42256-019-0103-7

Masten, C.L., Morelli, S.A., Eisenberger, N.I., 2011. An fMRI investigation of empathy for 'social pain' and subsequent prosocial behavior. NeuroImage 55 (1), 381–388, http://dx.doi.org/10.1016/j.neuroimage.2010.11.060.

Mischkowski, D., Crocker, J., & Way, B. M. (2019). A Social Analgesic? Acetaminophen (Paracetamol) Reduces Positive Empathy. *Frontiers in Psychology, 10*, 538-538. doi: 10.3389/fpsyg.2019.00538
McGinnis, D. What is the fourth industrial revolution? Salesforce https://www.salesforce.com/blog/2018/12/what-is-the-fourth-industrial-revolution-4IR.html (2018)

Mejía, S. T., & Hooker, K. (2017). Mixed emotions within the context of goal pursuit. *Current opinion in behavioral sciences*, *15*, 46-50.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*(4312), 75-78.

Meltzoff, A. N., & Moore, M. K. (1983). Newborn infants imitate adult facial gestures. *Child development*, 702-709.

Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental psychology*, *25*(6), 954.

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014). Bayesian inferences about the self (and others): a review. Conscious Cogn. 25, 67–76. doi: 10.1016/j.concog.2014.01.009

Newton, D. P., & Newton, L. D. (2019). Humanoid Robots as Teachers and a Proposed Code of Practice. *Frontiers in Education*, *4*, 125. https://doi.org/10.3389/feduc.2019.00125

Omohundro, Stephen M. 2008. "The Basic AI Drives." In Artificial General Intelligence 2008: Proceedings of the First AGI Conference,edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiersin Artificial Intelligence and Applications 171. Amsterdam: IOS

Parr, T., Rees, G., and Friston, K. J. (2018). Computational neuropsychology and Bayesian inference. Front. Hum. Neurosci. 12:61. doi: 10.3389/fnhum.2018.00061

Picard, R. W. (1997). *Affective computing*. MIT Press.

Prentice, C., Dominique Lopes, S., & Wang, X. (2020). Emotional intelligence or artificial intelligence–an employee perspective. *Journal of Hospitality Marketing & Management*, *29*(4), 377–403. https://doi.org/10.1080/19368623.2019.1647124

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *The Behavioral and Brain Sciences*, *25*(1), 1–20; discussion 20-71. https://doi.org/10.1017/S0140525X02000018

Rhue, L. (2018). Racial Influence on Automated Perceptions of Emotions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3281765

Robins, B., Dautenhahn, K., & Dickerson, P. (2009). From Isolation to Communication: A Case Study Evaluation of Robot Assisted Play for Children with Autism with a Minimally Expressive Humanoid Robot. *2009 Second International Conferences on Advances in Computer-Human Interactions*, 205–211. https://doi.org/10.1109/ACHI.2009.32

Schoeller F, Bertrand P, Gerry LJ, Jain A, Horowitz AH and Zenasni F (2019) Combining Virtual Reality and Biofeedback to Foster Empathic Abilities in Humans. Front. Psychol. 9:2741. doi: http://10.3389/fpsyg.2018.02741

Schoeller, F., Haar, A. J. H., Jain, A., & Maes, P. (2019). Enhancing human emotions with interoceptive technologies. Physics of Life Reviews. https://doi.org/10.1016/j.plrev.2019.10.008

Schoeller F, Miller M, Salomon R and Friston KJ (2021) Trust as Extended Control: Human-Machine Interactions as Active Inference. Front. Syst. Neurosci. 15:669810. doi: 10.3389/fnsys.2021.669810

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708). https://doi.org/10.1098/rstb.2016.0007

Shahina, A., Devosh, M., & Kamalakannan, N. (2014, April). EmoMeter: Measuring mixed emotions using weighted combinational model. In *2014 International Conference on Recent Trends in Information Technology* (pp. 1-6). IEEE.

Shuman, V., Sander, D., & Scherer, K. R. (2013). Levels of valence. *Frontiers in psychology*, *4*, 261

Smith, A. (2006). Cognitive empathy and emotional empathy in human behavior and evolution. *Psychological Record*, *56*(1), 3–21. https://doi.org/10.1007/BF03395534

Smith, A., & Anderson, M. (2017). *Automation in Everyday Life*. Pew Research Center. https://www.pewresearch.org/internet/2017/10/04/americans-attitudes-toward-robot-caregivers/

Soares, N., & Fallenstein, B. (2017). *Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda*. 103–125. https://doi.org/10.1007/978-3-662-54033-6_5

Sperduti, M., Guionnet, S., Fossati, P., Nadel, J. (2014) Mirror Neuron System and Mentalizing System connect during online social interaction, *Cognitive Processes, 15*: 307. https://doi.org/10.1007/s10339-014-0600-x

Spunt, R. P., & Lieberman, M. D. (2013). The Busy Social Brain: Evidence for Automaticity and Control in the Neural Systems Supporting Social Cognition and Action Understanding. *Psychological Science*, *24*(1), 80–86. https://doi.org/10.1177/0956797612450884

Sujan, Mark, et al. "Human factors challenges for the safe use of artificial intelligence in patient care." *BMJ health & care informatics* 26.1 (2019).

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational Identity: A Reader*, *56*(65), 9780203505984–16.

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2020). Alignment for advanced machine learning systems. *Ethics of Artificial Intelligence*, 342–382. https://doi.org/10.1093/oso/9780190905033.003.0013

Upshaw, M. B., Kaiser, C. R., & Sommerville, J. A. (2015). Parents' empathic perspective taking and altruistic behavior predicts infants' arousal to others' emotions. *Frontiers in Psychology*, *6*(APR), 1–11. https://doi.org/10.3389/fpsyg.2015.00360

Vaccaro, A. G., Kaplan, J. T., & Damasio, A. (2020). Bittersweet: The Neuroscience of Ambivalent Affect. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *15*(5), 1187–1199. https://doi.org/10.1177/1745691620927708

Vaish, A., Carpenter, M., & Tomasello, M. (2009). Sympathy Through Affective Perspective Taking and Its Relation to prosocial Behavior in Toddlers. *Developmental Psychology*, *45*(2), 534–543. https://doi.org/10.1037/a0014322

Williams, A., O'Driscoll, K., & Moore, C. (2014). The influence of empathic concern on prosocial behavior in children. *Frontiers in Psychology*, *5*(MAY), 1–8. https://doi.org/10.3389/fpsyg.2014.00425

Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do you feel my pain? Racial group membership modulates empathic neural responses. *Journal of Neuroscience*, *29*(26), 8525–8529.

Yampolskiy, Roman V., and M. S. Spellchecker. "Artificial intelligence safety and cybersecurity: A timeline of AI failures." *arXiv preprint arXiv:1610.07997* (2016).
van Geffen, W. H., Bruins, M., & Kerstjens, H. A. M. (2016). Diagnosing viral and bacterial respiratory infections in acute COPD exacerbations by an electronic nose: A pilot study. *Journal of Breath Research*, *10*(3), 036001. https://doi.org/10.1088/1752-7155/10/3/036001

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *IJCAI International Joint Conference on Artificial Intelligence*, *2018-July*, 5527–5533. https://doi.org/10.24963/ijcai.2018/779

Zaki, J., & Ochsner, K. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience*, *15*(5), 675–680. https://doi.org/10.1038/nn.3085

Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *PNAS*, 1–6.